자연어 처리 프로젝트

AI 챗봇 제작

김민섭
조영훈
김민경

목차

프로젝트 소개

여러가지 챗봇 사용해보기

다양한 챗봇 예제들을 실습해보고 이에 대한 결과를 비교해본다.

BERT 문장 임베딩 모델을 이용한 한국어 챗봇

SBERT를 이용하여 문장 임베딩을 얻을 수 있는 임베딩 모델을 사용하여 쉽고 간단하게 한국어 챗봇을 구현하는 예제
https://wikidocs.net/154530

트랜스포머를 이용한 한국어 챗봇

트랜스포머는 기존의 seq2seq의 구조인 인코더-디코더를 따르면서도, 어텐션(Attention)만으로 구현한 모델입니다.
이를 이용해 한국어 챗봇을 구현하는 예제 입니다.
https://wikidocs.net/89786

데이터 전처리 1. 원본 데이터

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		연령	성별	상황키워드	신체질환	감정_대분류	감정_소분류	사람문장1	시스템문장1	사람문장2	시스템문장2	사람문장3	시스템문장3
2	1	청년	여성	진로,취업,직장	해당없음	분노	노여워하는	일은 왜 해도 해도 끝(	많이 힘드시겠어요. 주	그냥 내가 해결하는 가	혼자 해결하기로 했군요. 혼자서 해결하기	힘들면 주위에 의논할	
3	2	청년	여성	진로,취업,직장	해당없음	분노	노여워하는	이번 달에 또 급여가	급여가 줄어 속상하시	최대한 지출을 억제해	월급이 줄어든 만큼 소비를 줄일 계획이군요.		
4	3	청년	여성	진로,취업,직장	해당없음	분노	노여워하는	회사에 신입이 들어왔	회사 동료 때문에 스트	잘 안 맞는 사람이랑	스트레스받지 않기 위해선 인간관계에 있어 약간의 거리를 두는		
5	4	청년	여성	진로,취업,직장	해당없음	분노	노여워하는	직장에서 막내라는 이	관련 없는 심부름을 5	직장 사람들과 술직하	직장 사람들과 이야기를 해 보겠다고 결심하셨군요.		
6	5	청년	여성	진로,취업,직장	해당없음	분노	노여워하는	얼마 전 입사한 신입사	무시하는 것 같은 태도	상사인 나에게 먼저 인	항상 먼저 인사하게 되어 화가 나셨군요. 어떻게 하면 신입사원에		
7	6	청년	여성	진로,취업,직장	해당없음	분노	노여워하는	직장에 다니고 있지만	진로에 대해서 고민하	직장 상사한테 자주 지	업무가 나와 맞지 않아 시간을 버리는 것 같이 느껴지셨군요.		
8	7	청년	여성	진로,취업,직장	해당없음	분노	노여워하는	성인인데도 진로를 아	부모님의 노여움에 섭	부모님께 좀 더 시간을	진로 테스트를 받아보겠다고 부모님께 말씀	드릴 생각이군요.	
9	8	청년	여성	진로,취업,직장	해당없음	기쁨	느긋	퇴사한 지 얼마 안 됐	천천히라도 직장을 구	더 좋은 회사가 기다리	좋은 일자리를 놓치고 싶지 않아 끊임없이 노력하시는군요. 하는		
10	9	청년	여성	진로,취업,직장	해당없음	불안	걱정스러운	졸업반이라서 취업을	취업에 대해 걱정이 되	응. 느긋함을 느끼는	걱정보다는 느긋한 태도가 낫다고 생각하시는군요.		
11	10	청년	여성	진로,취업,직장	해당없음	기쁨	느긋	요즘 직장생활이 너무	직장생활이 편하고 좋	우리 회사는 복지가 정	예비 엄마들에게는 정말 꿈의 회사인 것 같네요. 마음이 편하시겠		
12	11	청년	여성	진로,취업,직장	해당없음	기쁨	느긋	취업해야 할 나이인데	취업하고 싶지 않으시	아직 조금 더 놀고 싶	아직은 인생을 즐기고 싶은 마음이시군요.		
13	12	청년	여성	진로,취업,직장	해당없음	불안	당혹스러운	면접에서 부모님 직업	그때 어떤 생각이 들었	무척 당혹스러웠어. 부	그런 면접 상황에 무척 놀라셨군요. 어떻게 답하셨는지 여쭤봐도		
14	13	청년	여성	진로,취업,직장	해당없음	불안	당혹스러운	큰일이야. 부장님께 결	중요한 서류가 사라져	혹시 과장님이 가져가	혼이 날까 봐 과장님께 서류의 행방을 물어보는 것이 고민되시는		
15	14	청년	여성	진로,취업,직장	해당없음	불안	당혹스러운	나 얼마 전에 면접 본	회사에서 합격 통보받	코로나바이러스 때문(	차라리 면접을 보지 않았으면 더 좋았겠단 생각이 드시는군요.		
16	15	청년	여성	진로,취업,직장	해당없음	당황	당황	길을 가다가 우연히 ㅁ	취업에 대한 질문이 당	평소 친한 것도 아닌데	취직에 관해 물어보셔서 당혹스러운 상황이었군요.		
17	16	청년	여성	진로,취업,직장	해당없음	당황	당황	어제 합격 통보를 받은	잘못 발송된 문자로 당	물론 사과는 받았지. 대	대기업의 실수 때문에 실망감을 느끼시는군요.		
18	17	청년	여성	진로,취업,직장	해당없음	당황	당황	나 오늘 첫 출근 했는데	무슨 일 있었나요?	버스 타고 카드를 찍으	카드가 없어서 너무 당황스러웠겠어요.		
19	18	청년	여성	진로,취업,직장	해당없음	당황	당황	이번에 직장을 이직했	이직하신 직장에서 만	정말 다시는 만나고 싶	새 직장에서 만나고 싶지 않은 사람을 만나서 너무 스트레스를 받		
20	19	청년	여성	진로,취업,직장	해당없음	슬픔	마비된	코로나 때문에 뭘 할	코로나로 취업이 어렵	자격증 시험은 못 보지	시험은 못 보지만 공부는 계속하실 계획이시군요.		
21	20	청년	여성	진로,취업,직장	해당없음	슬픔	마비된	오늘 회사에서 큰 실수	무슨 실수를 하셨나요	중요한 발표가 있었는	준비 많이 하셨는데 힘드셨겠어요.		
22	21	청년	여성	진로,취업,직장	해당없음	슬픔	마비된	요즘 취업 관련해서 ㄸ	온몸에 마비될 것 같은	현재 취업 준비생인데	갈수록 힘들어지는 취업에 대해 생각하실 때마다 온몸이 마비된		
23	22	청년	여성	진로,취업,직장	해당없음	슬픔	마비된	어제도 야근 오늘도 0	야근이 많아서 힘드신	솔직히 말하자면 굳이	안 해도 될 야근을 하는 것 같아 더 스트레스받으시는 것 같군요.		
24	23	청년	여성	진로,취업,직장	해당없음	기쁨	만족스러운	우리 회사는 정말 사내	사내 분위기가 좋아서	즐거운 일이 매일 생길	원하는 진로에 맞춰서 취업해서 더욱 만족스럽군요.		
25	24	청년	여성	진로,취업,직장	해당없음	기쁨	만족스러운	오늘 내가 다니는 회사	직장이 좋으시군요. 지	회사에서 내가 제안한	존중하며 소통하는 직장에 크게 만족감을 느끼는군요.		
26	25	청년	여성	진로,취업,직장	해당없음	기쁨	만족스러운	회사에서 전공시험을	시험 결과가 어떠셨나	열심히 준비한 만큼 원	준비한 만큼 점수가 잘 나와서 만족스러우시겠어요.		
27	26	청년	여성	진로,취업,직장	해당없음	상처	배신당한	면접관에게 완전히 속	연봉과 실수령액이 차	회사 내규가 바뀌어서	신입 연봉체계가 다르시군요.		
28	27	청년	여성	진로,취업,직장	해당없음	상처	배신당한	지인이 취업시켜준다	지인에게 당한 사기라	취업시켜준다면서 십	믿었던 사람이라 배신감이 더 크시군요.		
29	28	청년	여성	진로,취업,직장	해당없음	상처	버려진	우리 부모님은 나를 ㅂ	부모님으로부터 서운	솔직히 말해보고 부모	부모님의 조언을 듣고 싶었는데 진중한 조언은 듣지 못했던 거군요		
30	29	청년	여성	진로,취업,직장	해당없음	상처	버려진	계속 취업이 안 되니까	계속 취업이 안 돼서	내가 능력이 없어서 취	취업 실패로 인해 자신감이 많이 떨어진 것 같네요. 힘내시길 바랄		
31	30	청년	여성	진로,취업,직장	해당없음	당황	부끄러운	이번에 중소기업에 취	중소기업에 취업해 부	요즘같이 어려울 때 취	오래 준비하고 공부해서 취업하셨군요.		
32	31	청년	여성	진로,취업,직장	해당없음	당황	부끄러운	저번 주에 친구와 같은	친구분과 같이 면접을	친구가 위로해주는데	친구의 위로에도 마음이 다스려지지 않는군요.		
33	32	청년	여성	진로,취업,직장	해당없음	분노	분노	요즘 청년 실업률이 ㄴ	청년 실업률의 심각성	기업들이 채용을 늘려	기업 채용을 늘려주는 정책이 생겼으면 좋겠다고 생각하시는군요.		
34	33	청년	여성	진로,취업,직장	해당없음	분노	분노	직장에서 부당한 일을	부당한 일을 당해서 ㅅ	이미 주변 사람에게 말	사람들이 공감해주지 않아서 더 화가 나는군요.		

데이터 전처리 2. Data shape 변경

```python
def set_data_shape(train_data):
  train_data1 = train_data[['사람문장1', '시스템문장1']]
  train_data2 = train_data[['사람문장2', '시스템문장2']]
  train_data3 = train_data[['사람문장3', '시스템문장3']]
  data_list = [train_data1, train_data2, train_data3]

  for i in range(len(data_list)):
    data_list[i] = data_list[i].rename(columns={f"사람문장{i+1}":"Q", f"시스템문장{i+1}":"A"})

  set_data = pd.concat(data_list)

  return set_data
```

```python
set_data_shape(train_data)
```

	Q	A
0	일은 왜 해도 해도 끝이 없을까? 화가 난다.	많이 힘드시겠어요. 주위에 의논할 상대가 있나요?
1	이번 달에 또 급여가 깎였어! 물가는 오르는데 월급만 자꾸 깎이니까 너무 화가 나.	급여가 줄어 속상하시겠어요. 월급이 줄어든 것을 어떻게 보완하실 건가요?
2	회사에 신입이 들어왔는데 말투가 거슬려. 그런 애를 매일 봐야 한다고 생각하니까 스...	회사 동료 때문에 스트레스를 많이 받는 것 같아요. 문제 해결을 위해 어떤 노력을 ...
3	직장에서 막내라는 이유로 나에게만 온갖 심부름을 시켜. 일도 많은 데 정말 분하고 ...	관련 없는 심부름을 모두 하게 되어서 노여우시군요. 어떤 것이 상황을 나아질 수 있...
4	얼마 전 입사한 신입사원이 나를 무시하는 것 같아서 너무 화가 나.	무시하는 것 같은 태도에 화가 나셨군요. 상대방의 어떤 행동이 그런 감정을 유발하는...
...	...	...
51628	함께 친하게 지내던 동네 언니 동생들과 빈자리를 조금이나마 채울까 해.	지인분들과 좋은 시간 보내셨으면 좋겠어요.
51629	사람들을 볼 때 의심하고 불신하는 마음을 억눌러야겠어. 사람들을 색안경을 끼고 보지...	원하시는 대로 가지고 계시던 걱정이 잘 해결되셨으면 좋겠어요.
51630	너 정말 똑똑하구나?	고마워요.
51631	좋은 밤 보내	안녕히 주무세요.
51632	큰 위로가 됐어	도움이 되었다니 기뻐요.

154899 rows × 2 columns

데이터 전처리 3.결측치 처리

결측치 처리

```
[ ]  train_data.isnull().sum()
```

```
Q    8935
A    8935
dtype: int64
```

```
train_data = train_data.dropna()
train_data
```

	Q	A
0	일은 왜 해도 해도 끝이 없을까? 화가 난다.	많이 힘드시겠어요. 주위에 의논할 상대가 있나요?
1	이번 달에 또 급여가 깎였어! 물가는 오르는데 월급만 자꾸 깎이니까 너무 화가 나.	급여가 줄어 속상하시겠어요. 월급이 줄어든 것을 어떻게 보완하실 건가요?
2	회사에 신입이 들어왔는데 말투가 거슬려. 그런 애를 매일 봐야 한다고 생각하니까 스...	회사 등료 때문에 스트레스를 많이 받는 것 같아요. 문제 해결을 위해 어떤 노력을 ...
3	직장에서 막내라는 이유로 나에게만 온갖 심부름을 시켜. 일도 많은 데 정말 분하고 ...	관련 없는 심부름을 모두 하게 되어서 노여우시군요. 어떤 것이 상황을 나아질 수 있...
4	얼마 전 입사한 신입사원이 나를 무시하는 것 같아서 너무 화가 나.	무시하는 것 같은 태도에 화가 나셨군요. 상대방의 어떤 행동이 그런 감정을 유발하는...
...	...	...
51628	함께 친하게 지내던 동네 언니 동생들과 빈자리를 조금이나마 채울까 해.	지인분들과 좋은 시간 보내셨으면 좋겠어요.
51629	사람들을 볼 때 의심하고 불신하는 마음을 억눌러야겠어. 사람들을 색안경을 끼고 보지...	원하시는 대로 가지고 계시던 걱정이 잘 해결되셨으면 좋겠어요.
51630	너 정말 똑똑하구나?	고마워요.
51631	좋은 밤 보내	안녕히 주무세요.
51632	큰 위로가 됐어	도움이 되었다니 기뻐요.

145963 rows × 2 columns

데이터 전처리 4. 정규표현식 적용

정규표현식 적용

```
train_data['Q'] = train_data['Q'].str.replace("[^가-힣?.!,]", " ")
train_data['Q'] = train_data['Q'].str.replace('^ +', "")
train_data
```

```
<ipython-input-15-866a6217a2b1>:1: FutureWarning: The default value of regex will change from True to False in a future version.
  train_data['Q'] = train_data['Q'].str.replace("[^가-힣?.!,]", " ")
<ipython-input-15-866a6217a2b1>:2: FutureWarning: The default value of regex will change from True to False in a future version.
  train_data['Q'] = train_data['Q'].str.replace('^ +', "")
```

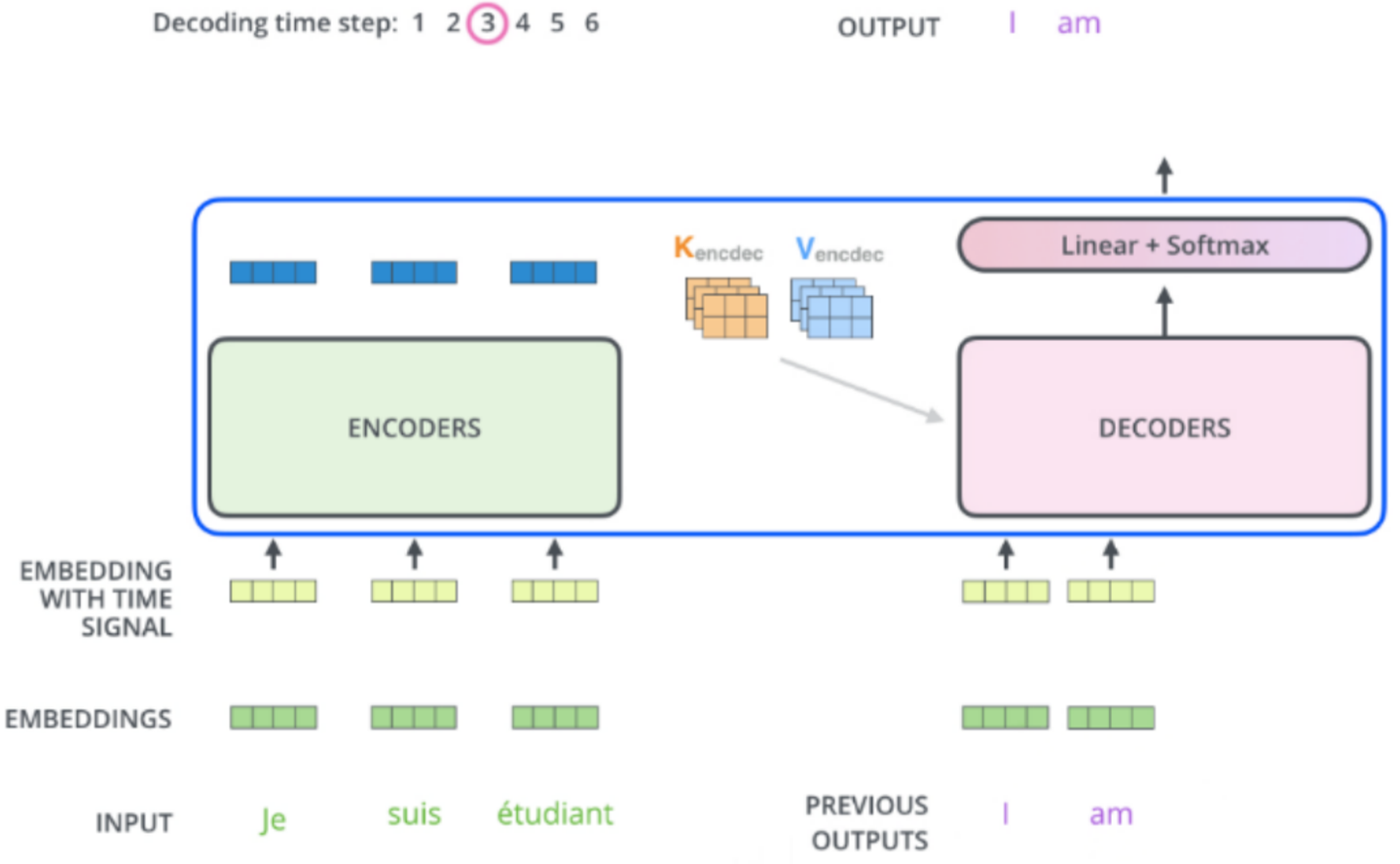
	Q	A
0	일은 왜 해도 해도 끝이 없을까? 화가 난다.	많이 힘드시겠어요. 주위에 의논할 상대가 있나요?
1	이번 달에 또 급여가 깎였어! 물가는 오르는데 월급만 자꾸 깎이니까 너무 화가 나.	급여가 줄어 속상하시겠어요. 월급이 줄어든 것을 어떻게 보완하실 건가요?
2	회사에 신입이 들어왔는데 말투가 거슬려. 그런 애를 매일 봐야 한다고 생각하니까 스...	회사 동료 때문에 스트레스를 많이 받는 것 같아요. 문제 해결을 위해 어떤 노력을 ...
3	직장에서 막내라는 이유로 나에게만 온갖 심부름을 시켜. 일도 많은 데 정말 분하고 ...	관련 없는 심부름을 모두 하게 되어서 노여우시군요. 어떤 것이 상황을 나아질 수 있...
4	얼마 전 입사한 신입사원이 나를 무시하는 것 같아서 너무 화가 나.	무시하는 것 같은 태도에 화가 나셨군요. 상대방의 어떤 행동이 그런 감정을 유발하는...
...	...	...
145958	함께 친하게 지내던 동네 언니 동생들과 빈자리를 조금이나마 채울까 해.	지인분들과 좋은 시간 보내셨으면 좋겠어요.
145959	사람들을 볼 때 의심하고 불신하는 마음을 억눌러야겠어. 사람들을 색안경을 끼고 보지...	원하시는 대로 가지고 계시던 걱정이 잘 해결되셨으면 좋겠어요.
145960	너 정말 똑똑하구나?	고마워요.
145961	좋은 밤 보내	안녕히 주무세요.
145962	큰 위로가 됐어	도움이 되었다니 기뻐요.

145963 rows × 2 columns

Bert

코드 리뷰

Transformer



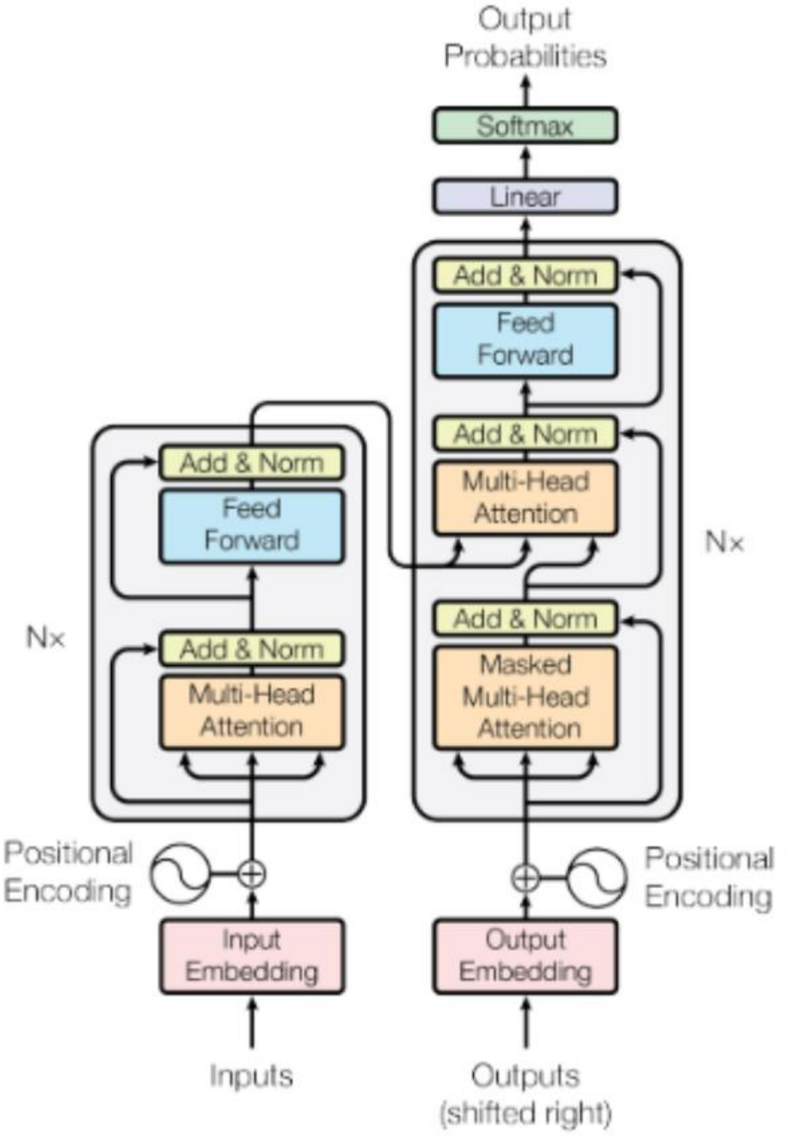


Figure 1: The Transformer - model architecture.

Transformer

데이터 수집 및 전처리

```
data = pd.read_csv('preprocessing_data.csv')
data
```

	Q	A
0	임은 왜 해도 해도 끝이 없을까? 화가 난다.	많이 힘드시겠어요. 주위에 의논할 상대가 있나요?
1	이번 달에 또 급여가 깎였어! 물가는 오르는데 월급만 자꾸 깎이니까 너무 화가 나.	급여가 줄어 속상하시겠어요. 월급이 줄어든 것을 어떻게 보완하실 건가요?
2	회사에 신입이 들어왔는데 말투가 거슬려. 그런 애랑 매일 봐야 한다고 생각하니까 스...	회사 동료 때문에 스트레스를 많이 받는 것 같아요. 문제 해결을 위해 어떤 노력을 ...
3	직장에서 막내라는 이유로 나에게만 온갖 심부름을 시켜. 일도 많은 데 정말 분하고 ...	관련 없는 심부름을 모두 하게 되어서 노여우시군요. 어떤 것이 상황을 나아질 수 있...
4	얼마 전 입사한 신입사원이 나를 무시하는 것 같아서 너무 화가 나.	무시하는 것 같은 태도에 화가 나셨군요. 상대방의 어떤 행동이 그런 감정을 유발하는...
...	...	...
145958	함께 진하게 지내던 동네 언니 동생들과 빈자리를 조금이나마 채울까 해.	지인분들과 좋은 시간 보내셨으면 좋겠어요.
145959	사람들을 볼 때 의심하고 불신하는 마음을 억눌러야겠어. 사람들을 색안경을 끼고 보지...	원하시는 대로 가지고 계시던 걱정이 잘 해결되셨으면 좋겠어요.
145960	너 정말 똑똑하구나?	고마워요.
145961	좋은 밤 보내	안녕히 주무세요.
145962	큰 위로가 되어	도움이 되었다니 기뻐요.

145963 rows × 2 columns

```
# 사용할 샘플의 최대 개수
MAX_SAMPLES = 145963
print(MAX_SAMPLES)
```

```
145963
```

데이터 수집

```
# 전처리 함수
def preprocess_sentence(sentence):
    # 양쪽의 공백을 제거합니다.
    sentence = sentence.strip()
    # 단어와 구두점(punctuation) 사이의 거리를 만듭니다.
    sentence = re.sub(r"([?.!,])", r" \1 ", sentence)
    # 여러 개의 공백을 하나의 공백으로 치환합니다.
    sentence = re.sub(r'[" "]+', " ", sentence)
    # (0-9 가-힣 ".", "?", "!", ",")를 제외한 모든 문자를 공백인 ' '로 대체합니다.
    sentence = re.sub(r"[^0-9가-힣?.!,]+", " ", sentence) #한글 전처리
    # 최종적으로 문장의 양쪽에 있는 공백를 제거합니다.
    sentence = sentence.strip()
    return sentence
```

```
# 질문과 답변의 쌍인 데이터셋을 구성하기 위한 데이터 로드 함수
# 데이터셋의 질문과 답변을 각각 inputs, outputs에 저장
def load_conversations():
    inputs, outputs = [], []

    for i in range(len(data)):
      # 전처리 후 리스트에 추가합니다.
        inputs.append(preprocess_sentence(data['Q'][i])) #questions
        outputs.append(preprocess_sentence(data['A'][i])) #answers

        if len(inputs) >= MAX_SAMPLES:  # 데이터 샘플의 최대 개수에 도달한 것으로 간주
            return inputs, outputs

    return inputs, outputs
```

전처리 후 질문과 답변 쌍 데이터셋 구성

Transformer

단어장 만들기

SubwordTextEncoder 사용

```
# 각 단어에 고유한 정수 인덱스를 부여하기 위해 질문과 답변 데이터셋에 대해서 Vocabulary 생성.
tokenizer = tfds.deprecated.text.SubwordTextEncoder.build_from_corpus(questions + answers, target_vocab_size=2**13)
# target_vocab_size=2**13: vocabulary의 크기를 2^13으로 설정

# 시작 토큰과 종료 토큰에 고유한 정수를 부여합니다.
START_TOKEN, END_TOKEN = [tokenizer.vocab_size], [tokenizer.vocab_size + 1]
```

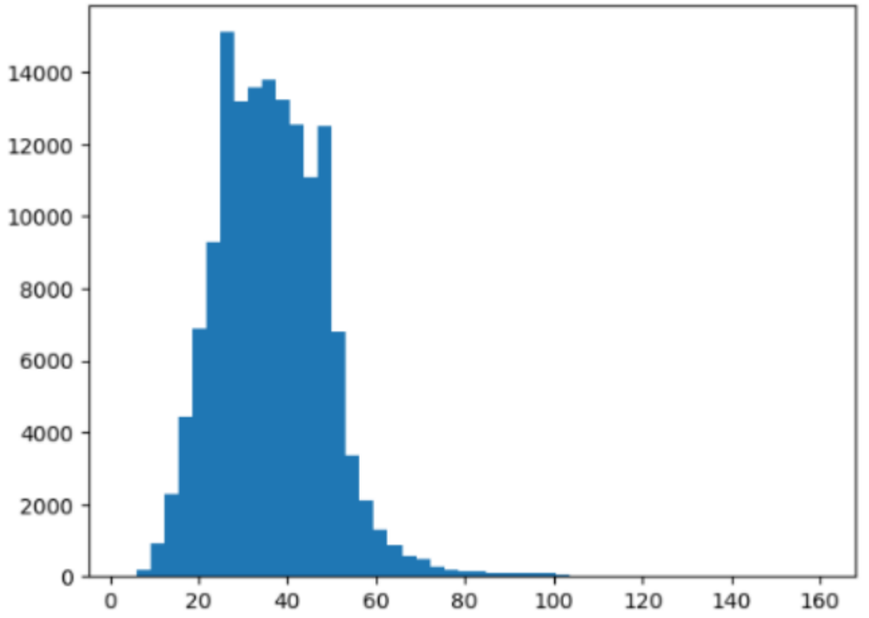
```
# 시작 토큰과 종료 토큰을 고려하여 +2를 하여 단어장의 크기를 산정합니다.
VOCAB_SIZE = tokenizer.vocab_size + 2
print(VOCAB_SIZE)
```

```
8240
```

Transformer

단어장 만들기

패딩 길이 설정



질문의 길이 시각화

```
# 전체 데이터에서 길이가 80이상인 질문의 비율 확인하기
cnt = 0
for q in questions:
    if len(q) > 80:
        cnt+=1
print(cnt)
```

```
758
```

```
cnt / 145963
# 0.519
```

```
0.005193096880716346
```

```
# 샘플의 최대 허용 길이 또는 패딩 후의 최종 길이
MAX_LENGTH = 80
print(MAX_LENGTH)
```

약 0.52% -> 제거

Transformer

단어장 만들기
정수 인코딩, 패딩, 교사강요

```python
def tokenize_and_filter(inputs, outputs):
    tokenized_inputs, tokenized_outputs = [], []

    for (sentence1, sentence2) in zip(inputs, outputs):
        # 정수 인코딩 과정에서 시작 토큰과 종료 토큰을 추가
        sentence1 = START_TOKEN + tokenizer.encode(sentence1) + END_TOKEN
        sentence2 = START_TOKEN + tokenizer.encode(sentence2) + END_TOKEN

        # 최대 길이 80 이하인 경우에만 데이터셋으로 허용
        if len(sentence1) <= MAX_LENGTH and len(sentence2) <= MAX_LENGTH:
            tokenized_inputs.append(sentence1)
            tokenized_outputs.append(sentence2)

    # 최대 길이 80으로 모든 데이터셋을 패딩
    # 입력 데이터셋 생성
    tokenized_inputs = tf.keras.preprocessing.sequence.pad_sequences(
        tokenized_inputs, maxlen=MAX_LENGTH, padding='post')
    # 출력 데이터셋 생성
    tokenized_outputs = tf.keras.preprocessing.sequence.pad_sequences(
        tokenized_outputs, maxlen=MAX_LENGTH, padding='post')

    return tokenized_inputs, tokenized_outputs
```

정수 인코딩, 패딩

```python
BATCH_SIZE = 128
BUFFER_SIZE = 20000  # 데이터셋을 섞기 위한 버퍼 크기

# 입력과 출력 데이터를 데이터셋으로 변환합니다.
dataset = tf.data.Dataset.from_tensor_slices((
    {
        'inputs': questions,
        'dec_inputs': answers[:, :-1] # 디코더의 입력, END_TOKEN이 제거된다.
    },
    {
        'outputs': answers[:, 1:]     # START_TOKEN이 제거된다.
    },
))

# 캐시를 사용하면 데이터를 파일에서 매번 다시 읽어오지 않고 메모리에 저장하여 처리 속도를 향상시킵니다.
dataset = dataset.cache()
# 데이터셋을 BUFFER_SIZE만큼 섞습니다.
dataset = dataset.shuffle(BUFFER_SIZE)
# 배치 크기만큼 데이터를 묶어서 처리합니다.
dataset = dataset.batch(BATCH_SIZE)
# 데이터를 미리 읽어서 메모리에 준비하도록 설정합니다.
dataset = dataset.prefetch(tf.data.experimental.AUTOTUNE)
# tf.data.experimental.AUTOTUNE: 자동으로 최적의 개수를 설정하여 데이터를 준비합니다
```

교사강요 사용

Transformer

모델 구성 및 평가

발전 방향 및 결론

감사합니다